

Technical Test

Big Data Cloud Developer

September 2024

Index

[1 Business Case 3](#_Toc178924733)

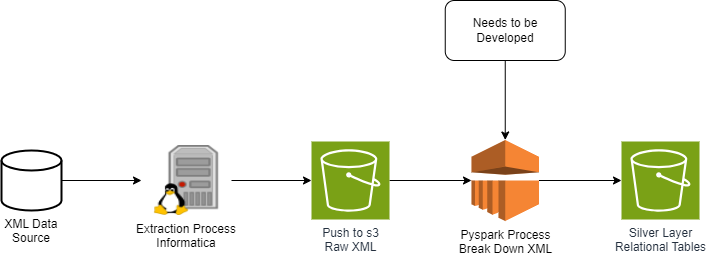
[2 Requirements 4](#_Toc178924734)

# Business Case

One of our customers wants to build analytical reports based on a XML source that they have internally. The BI users don’t know how to query the data in semi structured format and Management wants the data engineering team to break down the XML into relational tables. The architecture team provided the guidelines to use Pyspark for this task and a sample spreadsheet on how the tables should be relating to each other and how the PKs and FKs should be generated.

As this data will be fed later to a data mart, the output should be in parquet format (snappy).

This is the simplified architecture diagram of the solution provided by the Solutions Architect:



You need to develop the solution that will be running in the EMR cluster to break down the XML files.

For this task, please use Pyspark (any version).

This is the expected output and XML sample:



While coding, keep in mind the performance as this solution will be running against a large dataset, also. As an optional challenge, code in a way this solution could be reused for other business cases like this one.

Don’t worry about the integration with s3 and EMR now, you can work locally with Pyspark.

# Requirements

* The code should be in a repository Github (optional) or a zip file sent by email
  + Last commit must be 24 hours after receiving our email with the test. – don`t forget to set the repository as public
  + If you decide to send a zip file, your email must be sent 24 hours after receiving the test. Please keep in mind that mailboxes may block your attachment, therefore we strongly recommend publish it to github
* Create a README file containing any instructions you judge necessary
  + Example of this would be the version you use to code
  + Any specifics about your cod to run it (i.e. file must be in a specific path)
* Code in pySpark (recommended). If installing PySpark is a challenge, code in pure python is acceptable.